

COMPUTATIONAL SUBTRACTION METHOD

5

Field of the Invention

The invention relates to a method and system for detecting microbes harbored by a host organism. In particular, the invention relates to a method and system for detecting novel infectious disease organisms associated with the pathogenesis of human diseases.

Background of the Invention

10

Humans and animals are in continuous contact with microorganisms. Generally, because of the effectiveness of host defense mechanisms these microorganisms do not cause disease. However, some microorganisms (e.g., opportunistic pathogens) can become infective in particular types of individuals, such as those who are immunocompromised. Still other microorganisms are extremely virulent upon contact. For example, microorganisms such as the Ebola virus are associated with close to 100% fatality rates.

15

Traditional methods of correlating disease symptoms with the presence of a microorganism rely on identifying through symptoms and/or through epidemiological studies, the likelihood that the disease is caused by an infectious agent and attempting to culture appropriate samples from the individual to isolate and identify the agent. This can be problematic where epidemiological evidence is unclear, particularly in the case of pathogens with long incubation periods (e.g., up to 10 years in the case of HIV and 20-30 years in the case of Mycobacterium leprae).

20

25

Even where epidemiological evidence suggests an infectious cause for a disease, the microorganisms responsible for these diseases can evade detection. For example, Whipple's disease, a debilitating disease associated with diarrhea and weight loss, was for many years described as "intestinal lipodystrophy" because no microorganism could be cultured from samples from patients with disease. However, the microbial origin of Whipple's was suggested by the dramatic response of patients to antibiotics and the presence of bacilli observed in electron

micrographs of affected tissues. Still, the identification of the infectious organism as an actinomycetes awaited the advent of molecular techniques such as PCR. See, e.g., Maiwald et al., Clin. Infect. Dis. 32(3): 457-463 (2001). PCR amplification of conserved ribosomal sequences also led to the detection of another unculturable bacteria, the causative agent of bacillary angiomatosis which is associated with the proliferation of small blood vessels in the skin and visceral organs of patients with AIDS (see, Relman et al., New Engl. J. Med. 323: 1573-1580 (1990)). Another molecular technique, the DNA subtractive cloning method - representational difference analysis (Lisitsyn, Trends Genetics 11: 303-307 (1995)), enabled the discovery of the herpesvirus causing Kaposi's sarcoma (Chang et al., Science 265: 1865-1869 (1994)).

A high throughput approach to identifying infectious organisms has been described by Cummings and Relman, Emerg. Infect. Dis. 6(5): 513-25 (2000). Cummings and Relman report using a DNA microarray comprising sequences from known pathogens to detect the presence pathogens in patient samples. However, the method will only be able to detect pathogens for which at least some sequence information is known.

Summary of the Invention

There is a need in the art to provide a systematic approach for the detection and identification of microbes which are harbored within a host organism, particularly those associated with pathogenesis. Therefore, in one aspect, the invention provides a method of using a computer system to identify a microbe inhabiting a host organism which comprises the steps of obtaining sequence information from a plurality of sequences from at least one host organism and searching a database of host organism genomic sequences to determine the presence or absence of the plurality of expressed sequences in the database. The absence of at least one of the sequences in the database indicates that the at least one sequence is a candidate microbe sequence. Individual sequences can be searched sequentially; however, preferably, sets of sequences are searched at a time.

In one aspect, the method comprises the steps of obtaining sequence information from a library of genomic DNA from a host organism and searching a database of genomic sequences from host organisms to determine the presence or absence of a sequence in the library in the

database. A sequence that is present in the library but is absent in the database is identified as a candidate microbe sequence.

The microbe can be a symbiotic organism, such as a mutualistic organism, a commensal organism or a parasitic organism. The microbe can also be a pathogen. Microbes which can be identified by the method include, but are not limited to, phage, bacteria, viruses, protozoa and fungi. The host organism can be a microorganism, a plant, or an animal, such as a mammal (e.g., a human being). The host organism can also be an insect, bird, or a fish.

In one aspect, the plurality of sequences from the least one host organism comprises expressed sequences. For example, the plurality of sequences can comprise EST and/or cDNA sequences. Sequence information relating to expressed sequences can be obtained by sequencing a library of expressed sequences from one or more host organisms. Additionally, or alternatively, expressed sequence information can be obtained from a database of expressed sequences, such as an EST or cDNA database.

In one aspect, sequences from the at least one host organism suspected harboring a microbe are enriched for sequences which are present in the at least one host organism and which are not present in a plurality of host organisms which do not harbor the microbe. Enrichment can be performed using a subtractive hybridization assay, which can be a differential gene expression assay. Subtractive hybridization assays include, but are not limited to, representational difference analysis, SAGE, and suppression subtraction analysis. Alternatively, enrichment can be performed by electronically subtracting sequences from the at least one host organism which are stored in a first database from sequences of the plurality of organisms which are stored in a second database. In one aspect, the first and second databases are both expressed sequence databases and electronic subtraction is used to enrich for differentially expressed sequences which are expressed in the at least one host organism suspected of harboring a microbe and not expressed in the plurality of organisms which do not harbor the organism.

In one aspect, enriched sequences are then compared to sequences in a host organism genomic database to identify sequences in the at least one host organism suspected of harboring a microbe which are not present in the host organism genomic database. These sequences are identified as candidate sequences belonging to a microbe.

In a further aspect, one or more of the following sequences are eliminated from the host organism genomic database: vector sequences, mitochondrial sequences, repetitive sequences, sequences from other species, low quality sequences, known host organism mRNA sequences, and combinations thereof.

5 In a preferred aspect, the method according to the invention is used to identify the sequence of a pathogen. In this aspect, the at least one host organism is an organism which has a pathogenic condition, and sequences from the host organism (expressed or genomic) are compared to genomic sequences in a database from host organisms which do not have the pathogenic condition. The pathogenic condition can be a disease selected from the group
10 consisting of an inflammatory disease, an autoimmune disease, and a cell proliferative disease. More particularly, the disease can be selected from the group consisting of: sarcoidosis, inflammatory bowel disease (e.g., such as Crohn's disease), atherosclerosis, multiple sclerosis, rheumatoid arthritis, type I diabetes mellitus, lupus erythematosus, Hodgkin's disease, and bronchioalveolar carcinoma. Sequences from the at least one host organism which do not match
15 sequences in the genomic database identified as candidate sequences belonging to a pathogenic organism. In one embodiment, the pathogenic organism is an infectious disease organism.

In a further aspect, the invention provides a method of using a computer system to identify a microbe inhabiting a host organism, comprising the steps of: obtaining sequence information from a plurality of expressed sequences from at least one host organism; and
20 searching a database of host organism genomic sequences to determine the presence or absence of the plurality of expressed sequences in the database, wherein the absence of an expressed sequence in the database identifies the expressed sequence as a candidate microbe sequence. Preferably, the plurality of sequences are from a library. Still more preferably, the library is a library of expressed sequences. In one aspect, the library comprises human sequences. In
25 another aspect, the library comprises human sequences from one or more humans having a disease. The disease can be selected from the group consisting of an inflammatory disease, an autoimmune disease, and a cell proliferative disease. In one aspect, the disease is selected from the group consisting of sarcoidosis, inflammatory bowel disease, atherosclerosis, multiple sclerosis, rheumatoid arthritis, type I diabetes mellitus, lupus erythematosus, Hodgkin's disease,
30 and bronchioalveolar carcinoma.

In still a further aspect, the invention provides a method of using a computer system to identify a microbe inhabiting a host organism comprising the steps of: obtaining expressed sequence information from a plurality of sequences from at least one non-microbial host organism and searching a database of microbial sequences to determine the presence or absence of the plurality of expressed sequences in the database, wherein the presence of an expressed sequence in the database identifies the expressed sequence as a candidate microbe sequence. In one aspect, the plurality of sequences are from a library of expressed sequences. In another aspect, the library of sequences comprises sequences from one or more humans having a pathological condition, e.g., such as an infectious disease.

10 Candidate sequences can be used as query sequences to search a database of microbial sequences, such as a database comprising bacterial and/or viral sequences. Candidate sequences also can be used to search databases comprising fungal sequences, parasitic sequences, and/or protozoan sequences. Candidate sequences also can be used as query sequences to search a non-redundant expressed sequence database comprising sequences from host organisms.

15 Candidate sequences or their complements can be used to probe a library of sequences from at least one microbe to identify first hybridizing sequences, preferably sequences which are longer in length (e.g., numbers of bases) than the candidate sequence. Hybridizing sequences can in turn be used to identify second hybridizing sequences which are longer in length than the first hybridizing sequences. Overlapping sequences which are identified can be used to map the genomic structure of the microbe. In some aspects, the complement of the candidate sequence is hybridized to RNA from the microbe and used to generate cDNAs.

20 The candidate sequence can be used to express a peptide; for example, by operably linking the candidate sequence to a promoter sequence in an expression vector. Alternatively, or additionally, sequences identified by probing a library of sequences using the candidate sequence as a probe can be used to express one or more peptides. Preferably, the peptides are antigenic. Still more preferably, the peptides can be administered to a host organism to elicit a protective immune response. Nucleic acid sequences expressing the peptides can also be administered to the host organism to elicit a protective immune response to the peptides expressed by these sequences.

The candidate sequence and/or other sequences identified by the candidate sequence can be used to detect the presence or absence of the microbe in a sample from the host organism. For example, the hybridization of the candidate sequence and/or the other sequences to nucleic acid sequences in the sample from the host organism under stringent conditions can provide an indication of the presence of the microbe in the sample. In preferred embodiments, where the microbe is a pathogen, detection of hybridization is used to provide a diagnosis that the host organism is infected by the pathogen.

Peptides expressed by the candidate sequences and/or sequences identified using the candidate sequence can be used as antigens to generate antibodies which can also be used in diagnostic assays. For example, in one embodiment, an antibody which specifically binds to a peptide expressed by the candidate sequence and/or sequences identified using the candidate sequence is contacted with a sample from the host organism and binding of the antibody to a polypeptide within the sample provides an indication that the host organism harbors the microbe.

In some embodiments, the complementary sequence of a coding sequence of the candidate sequence or of another sequence identified by the candidate sequence is administered to a host organism harboring the microbe in an amount sufficient to prevent the expression of a polypeptide encoded by the candidate sequence or the sequence identified by the candidate sequence in the host organism. The complementary sequence can further comprise a cleaving moiety for cleaving RNA (e.g., the complementary sequence can be a ribozyme).

In one aspect, a system for performing the method is provided. The system comprises a first database comprising sequences from at least one host organism suspected of harboring a microbe and a second database comprising genomic sequences from host organisms not suspected of harboring the microbe. The system further comprises an information management system comprising a search and subtraction function for identifying sequences in the first database which are not present in the second database. In a preferred embodiment, the information management system comprises a sequence alignment function and can compare a set of sequences in the first database with all sequences in the second database. The system preferably comprises at least one user device connectable to the network and, preferably, a high speed, linear array processor.

In one aspect, the system comprises a program capable of implementing an algorithm for simultaneously comparing a plurality of sequences in a first database with all sequences in a second database, e.g., such as the algorithm implemented by the MEGABLAST program. However, in another aspect, the system comprises a program which sequentially compares a plurality of individual sequences from the first database with a plurality, and preferably all, sequences in the second database. Preferably, the system generates a result sequence set comprising sequences in the first database which do not match sequences in the genomic database.

In one aspect, the system comprises an identity or scoring matrix which requires a score of greater than or equal to 60 (e.g., equivalent to thirty identical consecutive nucleotides). In another aspect, the system iteratively computes the degree of alignment between sequences in the first and second database. Iterative computing preferably is performed using progressively smaller word sizes. In still a further aspect, the system provides one or more programs for performing one or more electronic subtraction functions for eliminating any of: vector sequences, repetitive sequences, mitochondrial sequences, sequences from non-host organisms, low quality sequences, known host organism mRNA sequences, and combinations thereof, from the genomic database.

The invention additionally provides a computer program product comprising a computer readable memory on which is embedded one or more programs for implementing any of the system functions and/or methods described above.

Brief Description of the Drawings

The objects and features of the invention can be better understood with reference to the following detailed description and accompanying drawings.

Figure 1 is a flow chart demonstrating a method of computational subtraction analysis according to one embodiment of the invention to identify microbes harbored by a human being.

Figure 2 is a schematic of a system according to one aspect of the invention for performing a computational subtraction analysis.

Detailed Description

The invention provides a method and system for performing computational subtraction to detect microbes harbored by a host organism. In some aspects, the microbes are pathogens and the system is used to identify sequences belonging to these pathogens which can then be used in methods of diagnosis and treatment. Alternatively, the microbes can be symbiotic organisms, such as commensal or parasitic organisms. Preferably, candidate sequences identified as belonging to a microbe are used to isolate and clone additional sequences from the microbe.

Definitions

The following definitions are provided for specific terms which are used in the following written description.

As used herein, the term "expressed sequence" is a sequence which is transcribed. "Expressed sequence information" refers to the nucleotide sequence of an expressed sequence such as an RNA molecule, a cDNA molecule or a portion of genomic DNA which corresponds to an expressed sequence, e.g., such as those portions of a gene whose complement will become part of an RNA transcript. An expressed sequence may include both coding sequences (i.e., codons which are translated into polypeptide sequences) as well as non-coding sequence (i.e., untranslated sequences).

As used herein, "a match" between sequences refers to a level of sequence similarity equivalent to a BLAST score ranging from 40 (the equivalent of 20 consecutive identical nucleotides) to 2000 (the equivalent of 1000 consecutive identical nucleotides)..

As used herein, a query sequence is "present" in a database if the database contains a sequence which matches the query sequence and is "absent" in a database if the database does not contain the matching sequence.

As used herein, a "low quality sequence" is a sequence which has greater than 2.5% N nucleotides, i.e., nucleotides whose identity cannot be determined at 95% confidence levels.

As used herein, "symbiosis" or a "symbiotic relationship" refers to an association between two organisms that live together. Symbiotic relationships include mutualistic relationships, commensalistic relationships, and parasitic relationships.

As used herein, "mutualism" or a "mutualistic relationship" refers to a mutually-
5 beneficial association between two organisms.

As used herein, "commensalism" or a "commensalistic relationship" refers to an association between two organisms where one organism may benefit but neither is harmed.

As used herein, "parasitism" or a "parasitic relationship" refers to an association between two organisms in which one organism lives at the expense of the other organism and can cause
10 damage to the other organism.

As used herein, a "pathogen" is an organism that can cause disease in another organism (e.g., the host organism).

As used herein, a "microbe" is any organism that can live and/or replicate within a host organism for at least a portion of its life cycle. While some microbes can exist for at least a
15 portion of their life cycle intracellularly within the cells of a host organism, microbes which grow and/or replicate extracellularly are also encompassed within the scope of the invention. Microbes include, but are not limited to, phage, viruses, gram-positive and gram-negative bacteria, protozoa, small unicellular and multicellular eukaryotes (e.g., fungi, such as yeast), and the like. The term "microbe" and "microorganism" are used interchangeably herein.

As used herein a "host organism" can any organism that can harbor (e.g., provide a habitat and/or nutrients for) another organism. Thus, a host can be a bacteria which harbors a phage, a simple eukaryote such as yeast which can harbor a bacteria, or a mammal such as a human being which can harbor by any of the foregoing.

As used herein, "infection" refers to the growth of a pathogen in a host organism.

As used herein, an "infectious disease" refers to a disease that can be transmitted from
25 host organism to host organism.

As used herein, a "carrier" refers to a patient who shows full recovery after infection and displaying symptoms but still carries and is capable of spreading the infectious form of a microbe.

As used herein, "a sequence identified by a candidate sequence" refers to genomic
5 sequences of microbes to which the candidate sequence or its complement hybridizes, or to which the latter genomic sequences hybridize, under stringent conditions. In some embodiments, sequences are identified by the candidate sequence electronically, e.g., by searching a database of sequences from one or more microbes. Sequences which are identified as belonging to the same microbe as the organism from which the candidate sequence was
10 obtained are said to be "identified by the candidate sequence."

As used herein, "stringent conditions" refer to conditions under which a sequence will specifically bind to its complement to enable detection of the complement and to distinguish the complement from other nucleic acid sequences in a sample. Stringency conditions are described
15 in Sambrook et al., In Molecular Cloning: A Laboratory Manual, 2nd edition, vols. 1-2, Cold Spring Harbor Press (1989), the entirety of which is incorporated by reference herein. As used herein, stringent conditions require at least 80% base pairing, more preferably, at least 90-95% base pairing, and most preferably, at least 98% base pairing.

As used herein, a "fragment" of a candidate sequence or a sequence identified by the candidate sequence refers to a sequence which is shorter in length than the candidate sequence
20 but sufficiently long to specifically hybridize to the candidate sequence. In one embodiment, a fragment ranges in size from 6 nucleotides to one less nucleotide than the full-length sequence.

As defined herein, the "a promoter operably linked" to another sequence refers to a promoter and/or promoter element and/or enhancer element(s) capable of inducibly or constitutively causing transcription of the other sequence.

25 As used herein, a "bodily fluid" refers to any of blood, plasma, sera, urine, CSF fluid, sputum, breast exudates, pus, and the like.

As used herein, "computational subtraction" or "electronic subtraction" or "filtering" refers to a computational method of eliminating records (e.g., such as sequences) from a database.

Computational Subtraction

5 In one aspect, the invention provides a systematic method to identify sequences of microbes capable of inhabiting a host organism. The microbes can be pathogenic and associated with an infectious disease. However, the microbes can also exist symbiotically within a host organism, e.g., in a mutualistic, commensal, or parasitic relationship within the host organism. The microbe can be any of a phage, a virus (e.g., an RNA or DNA virus), a bacteria, a protozoa,
10 or other microorganism, a small unicellular or multicellular eukaryotic organism (e.g., a fungus, such as yeast), and the like. The host organism can be a microorganism, a fungus, an animal, or a plant. Preferably, the host organism is a mammal, such as a human being or a domestic animal. However, the host organism can also be an insect, bird, or a fish. Host organism sequences can be obtained from particular tissues or cells of the host organism, or from cell lines derived from
15 these tissues or cells, or from bodily fluids from the host organism.

The invention provides a computational subtraction method for detecting and identifying microbe sequences. The method comprises comparing the sequence information of a plurality of sequences obtained from one or more host organisms with sequences in a genomic database of host sequences to identify which of the plurality of sequences are not found (i.e., do not match
20 other sequences) in the database. Sequences which are not found in the database are identified as candidate sequences which are likely to belong to a microbe. Preferably, sequence information from sets of sequences (two or more sequences, and preferably ten or more sequences) are compared against the entire genomic database at a time.

Any number of nucleotide sequence alignment algorithms can be used for this purpose,
25 including those known in the art. For example, in one aspect, the algorithm of Zhang et al., J. Comput. Biol. 7(1-2): 203-14 (2000), which is embodied in one form in the MEGABLAST program, is used to compare sequences in an entire database of sequences from one or more host organisms (a "test database") against a genomic database. Smaller sets of sequences (e.g., at least two or at least ten) can also be compared. In some aspects, sequences from the plurality of

sequences can be compared sequentially, individually, against genomic databases, e.g., such as by using the BLAST program described in Altschul et al., J. Mol. Biol. 215: 403-410 (1990); Madden et al., Meth. Enzymol. 266: 131-141 (1996); and Zhang et al., Genome Res. 7: 649-656 (1997), the entireties of which are incorporated by reference herein. Other programs whose goal is sequence similarity searching also can be used, such as FASTA, SSAHA, or any type of word hashing program such as is known in the art (see, e.g., Pearson, Proc. Natl. Acad. Sci. USA 85(5): 2444-2448 (1988); Leung et al., J. Mol. Biol. 221(4): 1367-1378 (1991), the entireties of which are incorporated by reference herein).

Methods of determining the significance of sequence alignments are known in the art and are described in Needleman and Wunsch, J. of Mol. Biol. 48: 444 (1970); Waterman et al., J. Mol. Biol. 147: 195-197 (1980); Karlin et al., Proc. Natl. Acad. Sci. USA 87: 2264-2268 (1990); Karlin et al., Proc. Natl. Acad. Sci. USA 90: 5873-5877 (1993); Dembo et al., Ann. Prob. 22: 2022-2039 (1994) and Altschul, In Theoretical and Computational Methods in Genome Research. (S. Suhai, ed.), pp. 1-14, Plenum, New York; the entireties of which are incorporated by reference herein.

In some aspects, the genomic database is searched for short perfect matches of a set length (i.e., a word size). This enables a more rapid comparison than window/stringency matching. In one embodiment, a word size ranging from 10-30 bases is used. Preferably, a series of sequential searches is performed, using progressively smaller word sizes ranging from 30 to 10 bases. More preferably, a first search using a word search of 24 is performed, followed by a second word search of 20, followed by a third word search of 16, followed by a fourth word search of 12. In one aspect, a test sequence is shifted to the left or right of sequences in the database to identify maximal regions of alignment.

In some aspects, a scoring matrix is used to identify the likelihood that one or more sequences in the test database do not match or are absent from the genomic database. Preferably, scores of greater than or equal to 60 are required. In one aspect, the scoring matrix assigns a match if there is a BLAST score ranging from 40 (the equivalent of 20 consecutive nucleotides) through 2000 (the equivalent of 1000 consecutive nucleotides). In another aspect, a matrix is used which assigns expectation values to matches and mismatches after alignment. Expectation

values can be adjusted to require that a score does not grow simply by extending the alignment in a random way. For example, in one embodiment, expectation values of from 10^{-20} – 10^{-3} can be selected, and preferably, expectation values of 10^{-7} are used. Gap values can be set to any desired value as is routine in the art (see, e.g., Smith et al., 1981, J. Mol. Evol. 18(1): 38-46, Levitt et al., 1998, Proc. Natl. Acad. Sci. USA 95(11): 5913-5920, the entireties of which are incorporated herein by reference.

In one aspect, a results database is created, preferably comprising sequences from the test database which are ranked according to their alignment with sequences in the genomic database. Preferably, sequences which show a high degree of alignment to genomic sequences from the host organism (e.g., having at least 20-1000 consecutive identical nucleotides) are not included in the results database or are subsequently removed from the results database.

In a preferred embodiment, as shown in Figure 1, a subtraction operation is performed to remove sequences from either the genomic database and/or the test database and/or the results database. For example, subtraction operations can be used to remove vector sequences, repetitive sequences, mitochondrial sequences, sequences from other species, low quality sequences, known host organism mRNA sequences, and the like. It should be obvious to those of skill in the art that the order of subtraction operations is not critical and that one or more subtraction operations can be used. In certain aspects, after filtering operations to filter sets of candidate sequences through one or more of a vector sequence database, a repetitive sequence database, a mitochondrial sequence database, a non-host species database, and/or a known host organism mRNA database ("filtering databases"), a first candidate sequence set of sequences is again compared to the host organism genomic database, and/or one or more filtering databases using a reduced word size than was used in the previous series of operations, to generate a second candidate sequence set which is then stored in a results database. In a preferred aspect, low quality sequences are removed, before or after filtering.

Test Database

In one aspect, the test database is an expressed sequence database of sequences from the host organism, such as an EST or cDNA database (e.g., a library database). Such databases are known in the art and include, but are not limited to, human expressed sequence databases such as

the NCBI EST database, the LIFESEQTM database (Incyte Pharmaceuticals, Palo Alto, Calif.), the random cDNA sequence database from Human Genome Sciences, the EMEST8 database (EMBL, Heidelberg, Germany), and the like (see, also, Boguski et al., 1993, Nat. Genet. 4(4): 332-333, the entirety of which is incorporated by reference herein).

5 The test database also can be generated by inputting and storing sequence information obtained by sequencing a plurality of nucleic acids from a library of expressed sequences from one or more host organisms suspected of harboring a microbe, into a user device of a system 1 (shown in Figure 2) as described further below. Libraries of expressed sequences can be generated using total RNA or polyadenylated RNA, and by using random priming or oligodT
10 priming or a combination of these methods. Such techniques are known in the art. Libraries of particular interest include, but are not limited to, libraries of expressed sequences from one or more patients with an inflammatory disease, an autoimmune disease, and a cell proliferative disease. For example, in one aspect, libraries of expressed sequences from one or more patients with a disease selected from the group consisting of sarcoidosis, inflammatory bowel disease
15 (such as Crohn's disease), atherosclerosis, multiple sclerosis, rheumatoid arthritis, type I diabetes mellitus, lupus erythematosus, Hodgkin's disease, and bronchioalveolar carcinoma are used to obtain expressed sequence information. Preferably, the creation of such libraries is performed to minimize manipulation of tissue (e.g., by careful attention to sterility and avoidance of
20 amplification methods) to avoid spurious contamination of such libraries with bacterial sequences.

 While the test database can consist of entirely expressed sequences, the test database can also be a genomic sequence database. For example, the test database can comprise sequence information from a plurality of sequences in a genomic library from one or more host organisms suspected of harboring a microbe. Preferably, genomic sequence test databases are used to
25 identify expressed sequences of microbes which are not polyadenylated (and/or which have integrated into the genome of the host organism), e.g., such as bacterial expressed sequences which would likely escape detection in expressed sequence libraries generated from polyadenylated RNA.

The test database can be enriched for sequences which are found in host organism(s) suspected of harboring a microbe and which are not found in host organisms not harboring the microbe. In one aspect, the enrichment method comprises combining genomic test sequences with other genomic sequences (reference sequences), expressed test sequences with expressed reference sequences, or expressed test sequences with genomic reference sequences, and removing sequences which are common to both test and reference sequence sets, thereby enriching for test sequences which are not found in a reference set of sequences.

For example, in one aspect, a subtractive hybridization method is used to enrich for expressed sequences in a sample of nucleic acids from a host organism which is suspected of harboring a microbe and which are not expressed in host organisms which do not harbor the microbe. Samples can comprise total nucleic acids, polyadenylated RNA, or total RNA. Subtractive hybridization methods to enrich for differentially expressed sequences are known in the art and include, but are not limited to, SAGE (Serial Amplification of Gene Expression) (see, e.g., Velculescu et al., *Science* 270: 484 (1995) and U.S. Patent No. 5,866,330), subtractive hybridization of cDNA libraries (e.g., using magnetic beads, as described in WO 97/07244 A1), cDNA representational difference analysis (e.g., Hubank and Schatz, *Nucl. Acids Res.* 22: 5640-5648 (1994)), and suppression subtraction analysis (see, e.g., U.S. Patent 5,565,340). Subtractive hybridization methods can also be used to enrich for sequences which are present at different levels in different populations of genomic DNA. Such methods include, but are not limited to, representational difference analysis, such as described in U.S. Patent No. 5,958,738 and CLONTECH's PCR-Select™ Bacterial Genome Subtraction technique (see, e.g., Diatchenko et al., *Proc. Natl. Acad. Sci. USA* 93: 6025-6030 (1996); CLONTECHniques X(4): 2-5 (1995)). The entireties of these references are incorporated by reference herein.

In another aspect, enrichment is performed electronically. For example, sequences from at least one host organism suspected of harboring a microbe stored in a test database can be subtracted from sequences in a "reference database" comprising sequences from a plurality of host organisms not harboring the microbe. In one aspect, the test database and reference database are both expressed sequence databases and electronic subtraction is used to enrich for differentially expressed sequences which are expressed in the at least one host organism and which are not expressed in the plurality of host organisms. Methods for electronic subtraction

analysis of expressed sequences are described in U.S. Patent 6,114,114, for example, the entirety of which is incorporated by reference herein.

In some embodiments, the test database is a relational database which segregates particular types of sequences from other types of sequences within the database. For example, in one aspect, expressed sequence information can be subdivided within an expressed sequence database according to a particular tissue, or cell type, or cell line, in which the sequence is expressed. In these embodiments, particular portions of the test database can be compared to the genomic database during a search, sequentially, or simultaneously.

In one aspect, once a candidate sequence is identified, it is compared to a nucleotide sequence database comprising sequences from a plurality of species, to identify the microbial organism genus to which the sequence belongs or to which the species is related evolutionarily. For example, GenBank's nucleotide or "nt" database can be used. In another aspect, the candidate sequence can be used as a query sequence to search a database comprising only microbe sequences. In one embodiment, the database is a microbial sequence database which can be a viral sequence database, or a fungal or parasite sequence database. Such databases are known in the art and include, but are not limited to, the Incyte Microbial Database, the TIGR Microbial Database, the TIGR Parasites Database, TIGR Fungal Database, and the TIGR Viral Genome Sequencing Project Database. This step can be used to identify or evaluate the taxonomic relationship between the candidate sequence and sequences of other known microbes for which genomic sequence information is known.

In still another aspect, candidate sequences are compared to sequences in a non-redundant RNA database to determine whether the sequence matches known host organism RNA molecules. In still a further aspect, a candidate sequence is conceptually translated to identify open reading frames and the amino acid sequences of a polypeptide encoded by the candidate sequence can be used as a query sequence to search a protein sequence database comprising microbial sequences (i.e., the database can comprise multiple species sequences in addition to microbial sequences, such as the GenBank nr database, or the database can comprise exclusively microbial sequences). Preferably, the sequence is also used as a query sequence to search a nucleotide sequence database comprising microbial sequences (e.g., such as the GenBank nt

database or an exclusively microbial sequence database) to identify sequence whose conceptual translations match known microbial proteins but whose nucleotide sequences do not match microbial nucleotides. These latter classes of sequences, which are preferably stored in a results database, are likely to identify sequences belonging to microbes of the same genus as the
5 microbe whose protein was identified as a match, but which do not necessarily represent microbes belonging to the same species, i.e., the sequences are likely to represent previously uncharacterized microbes.

Genomic Databases

Genomic databases for a variety of host organisms are also known in the art, and include,
10 but are not limited to, the NCBI GenBank database (see, e.g., <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome>), the Celera Human Genome (<http://www.celera.com>), the Genetic Information Research Institute (GIRI) (<http://www.girinst.org>) and Human Genome Fragment database, TIGR databases (e.g., the TIGR Human Gene Index Database), and the like.

The genomic database also can be generated by inputting and storing sequence
5 information obtained by sequencing a plurality of nucleic acids from a genomic library of sequences from one or more host organisms which do not harbor microbes, into a user device of a system described further below.

Genomic databases contemplated according to the invention include genomic sequence information from any of the host organisms described above, e.g., from a microorganism, a
20 fungus (e.g., yeast), an animal (insect, bird, fish, or mammal, such as a human being or domestic animal) or a plant.

System For Performing The Computational Subtraction Method

The invention further provides a system 1 for performing the computational subtraction analysis described above (see, Figure 2). In one aspect, the system 1 comprises a first database 2
25 (e.g., the test database) comprising sequences from at least one host organism suspected of harboring a microbe and a second database 3 comprising genomic sequences from host organisms not harboring the microbe. The system 1 further comprises an information management system 4 comprising a search function for identifying sequences in the first

database 2 which are not present in the second database 3. In a preferred embodiment, the system 1 further comprises a program embodied in a computer readable medium for executing sequence alignments between at least a first sequence in the first database and a plurality, and preferably, all sequences in the second database. The program can be part of a server 5 (which also can store program applications required by the information management system 4) or part of a processor which is part of a user device 9. Preferably, however, the user device 9 is in communication with the server 5 and/or other servers (not shown). The user device 9 can be a computer, a laptop, a wireless device, and the like. The system further can include additional user devices 9, output devices 6 (e.g., printers), and input devices (e.g., keyboards 7, mice, joysticks, and the like). The user device 9 preferably includes an interface 8 which can be displayed by the device 3 in response to the user accessing the system 1 to activate the information management system 5. The system 1 is preferably connectable to the network 10, enabling a user to access the system remotely from any user device 3 that is connectable to the network 10.

In preferred embodiments, sets of sequences (at least 2, at least 10, at least 100, or at least 500) in the test database 2 are compared at a single time with sequences in the genomic database 3. In one embodiment, the information management system 5 comprises a program which is capable of implementing an algorithm, such as the one used in the MEGABLAST program for performing this function (see, e.g., Zhang et al., supra). However, in other embodiments, the at least 2, at least 10, at least 100, or at least 500 sequences are compared individually and sequentially with sequences in the genomic database 3.

In preferred embodiments, the user device 3 or the server 5 (or another host computer) comprises a high speed, linear array processor that can locate highly similar sequence segments (e.g., having a BLAST score of at least 40) from any at least two sequences. In one aspect, the processor comprises a high speed circuit chip that provides an equivalent of about 400,000 transistors or 100,000 gates, as described in U.S. Patent No. 5,964,860, for use in performing high speed sequence analyses.

In one aspect, the system 1 further comprises an input device 7 that receives a set of sequences (either sequentially or simultaneously), a memory that stores the set of sequences (not

shown), and a processor that transfers information from the set of sequences to the memory (e.g., in the form of data characters representing nucleotide bases in the set of sequences). The processor can be part of a user device 3, but is preferably part of the server 5. In another aspect, the system 1 further includes an identity matrix and a result sequence set (e.g., from the results database described above) (not shown), in which members of a set of compared sequences are ranked according to their degree of match to sequences in the genomic database 3. In a further aspect, the results sequence set can include sequences which do not match sequences in the genomic database. Sequences which have the least amount of match (as determined using parameters established by the user) can be displayed on an interface 8 of the user device 8 in response to a user query to match sequences.

In one aspect, the identity matrix is pre-selected by the user to require a match score of greater than or equal to 60 with a word size of between 10 and 30. In one embodiment, the system 1 iteratively computes the degree of alignment between sequences using progressively smaller word sizes from 30 to 10, (e.g., first using a word size of 24, then a word size of 20, then a word size of 16, then a word size of 12). Preferably, the score value remains the same and is some value greater than or equal to 60. The matrix is designed to eliminate low quality sequences (e.g., as determined using a base calling program such as PHRED), short sequences (less than 150 nucleotides), or sequences comprising a maximum number of ambiguous or unreadable nucleotides, such that there is a minimum length of quality sequences (e.g., sequences whose bases have a high confidence (at least 95%) of being accurate) of at least 50 nucleotides, and preferably at least 150 nucleotides.

In one aspect, the system 1 provide one or more programs for performing one or more electronic subtraction functions analogous to an electronic subtractive hybridization. For example, in one aspect, the system 1 is capable of eliminating, in response to a user command or in response to a pre-programmed set of instructions, any of: vector sequences, repetitive sequences, mitochondrial sequences, sequences from other species, low quality sequences, known host mRNA sequences (i.e., sequences known to belong to the host organism), and combinations thereof.

In a further aspect, the invention provides a computer program product comprising a computer readable memory on which is embedded one or more programs for implementing any of the system 1 functions described above.

Methods of Using Candidate Sequences

5 Cloning and Sequencing Genomic DNA From Microbes

In one aspect, candidate sequences identified using the methods and system 1 described above are used as probes to probe a library of sequences from at least one microbe. The microbe can be a phage, a virus, a bacteria, a protozoa, or other microorganism, a small unicellular or multicellular eukaryotic organism, such as a fungi (e.g., yeast), and the like. Microbes can be
10 cultured from host organisms to provide nucleic acids suitable for generating libraries using methods known in the art. Preferably, the library is a genomic library. Alternatively, where microbes cannot be cultured, libraries can be generated from genomic or expressed sequences from which host organism sequences have been subtracted as described above. In some aspects, microbe sequences which are enriched in these samples can be ligated to linkers or adapters and
15 amplified using primers which hybridize to these linkers or adapters. Alternatively, or additionally, the linkers or adapters can include promoter sequences and microbe sequences can be amplified by providing polymerases which recognize these sequences and the appropriate nucleotides (e.g., using a transcription-based amplification system). These methods can be complemented by additional rounds of computational subtraction as described above by
20 sequencing enriched sequences and subtracting sequence information corresponding to these enriched sequences from a genomic database to identify enriched sequences which are not found in the genomic database.

In one aspect, candidate sequences are used to identify hybridizing sequences within the library which are longer in length than the candidate sequence, either at the 5' end or 3' end or
25 both. These longer sequences are used, in turn, to identify other sequences which are preferably longer in length either at the 5' end or 3' end or both. Overlapping clones can be mapped using restriction enzyme analysis in combination with Southern analysis, and/or sequence analysis, to further characterize the genome structure of the microbe. Preferably, genomic sequence

information is inputted into a microbial genomic database (i.e., a database comprising only microbial sequences).

Microbe sequences can be evaluated using a sequence analysis program such as the Gene Locator and Interpolated Markov Modeler, or GlimmerTM, program to identify coding sequences and to distinguish such sequences from non-coding DNA (see, e.g., Salzberg et al. Nucl. Acids Res. 26(2):544-8 (1998). A version of Glimmer designed for small eukaryotes is described in Salzberg et al., Genomics 59: 24-31 (1999). The entirety of these references is incorporated by reference herein.

In one embodiment, RNA samples are obtained from host organisms harboring the microbe (e.g., total RNA or polyA RNA if the microbe's RNA is polyadenylated) and a complement of the candidate sequence is used as a primer to generate cDNA molecules from the RNAs obtained. In one embodiment, cDNAs are generated using a RACE method (see, e.g., Siebert et al., In Gene Cloning and Analysis by RT-PCR (BioTechniques Books, Natick, MA), pp. 305-320 (1998); Don et al., Nucl. Acids Res. 19: 4008 (1991); Roux, PCR Methods Appl. 4: 5185-5194 (1995); the entireties of which are incorporated by reference herein) to identify the 5' and or 3' end of a particular RNA transcript. Preferably, the sequences of cDNA clones are inputted into a results database for comparison to a database comprising microbial nucleotide sequences.

In some aspects, when a host organism has been identified, candidate sequences and/or their complements can be used as primers in PCR or RT-PCR assays to identify additional microbial sequences of interest, for example, in nucleic acids obtained from cultures of these microbes. In one aspect, asymmetric or one-directional PCR can be performed using the candidate sequence or its complement as a single primer in primer extension reactions to identify microbial sequences flanking the primer sequence in the microbial genome or in a microbial transcript. One-directional PCR is known in the art and is described in U.S. Patent No. 6,184,025, for example, the entirety of which is incorporated by reference. In other aspects, at least two primers corresponding to the candidate sequence are used (e.g., primers capable of amplifying a nucleic acid fragment which comprises a subsequence of the candidate sequence of at least 50 nucleotides). In addition to detecting microbial sequences amplifiable using these

primers, the primers can be used to verify that the candidate sequences do not represent previously unsequenced host genomic DNA. For example, the primers can be used in amplification reactions with host genomic DNA to verify that no amplification of host genomic sequences occurs.

5 Diagnostic Methods

Candidate sequences, their complements, or sequences identified by candidate sequences (e.g., such as by any of the assays described in the preceding section), can be used in hybridization assays to detect the presence of a microbe in a sample. Although such methods are described as "diagnostic", this does not imply that the method is necessarily used to determine the presence or absence of a pathogenic condition in an organism. For example, diagnostic methods can be used to detect the presence of a commensal microbe within a sample, which can, in some instances, be desirable (e.g., such as when the microbe produces vitamins for the host). In some instances, however, the hybridization assays can be used to detect the presence of one or more pathogens in a sample from an organism, and the results of such as assay can be used to provide treatment options for the organism. In still other aspects, the hybridization assays are used to detect carrier organisms which are infected by pathogens but which do not show symptoms of a pathogenic condition.

In one aspect, nucleic acids from a sample obtained from a host organism (e.g., a cell, a tissue sample, a bodily fluid, a lavage specimen, and the like) are contacted under stringent conditions with a test sequence derived from the candidate sequence. As used herein, "a test sequence derived from a candidate sequence" refers to the candidate sequence itself, or a fragment thereof, or another sequence from the microbe which the candidate sequence has been used to identify, or to complements of any of these sequences.

The test sequence can be used as a diagnostic probe to detect expressed sequences or genomic sequences of the microbes in the sample by detecting the formation of a hybridization complex between the test sequence and a nucleic acid in the sample. In one embodiment, test sequences are labeled with detectable labels. However, in other embodiments, the test sequence is bound to a molecule which is detectably labeled or which itself can bind to detectably labeled molecule(s). In one aspect, the amount of test sequences bound is used to provide an indication

of the number of microbes in a sample (for example, by providing a comparison to test samples comprising a known amount of microbes). In another aspect, either the sample sequences, or probe sequences, or both, are amplified (e.g., by PCR, LCR or some other means of amplification) to increase the sensitivity of the assay. In still a further embodiment, the test sequence itself is used as a primer in an amplification assay or a reverse transcription-based assay. Methods of labeling, hybridizing, amplifying and quantitating nucleic acids are known in the art. Probes can be obtained by restriction digestion of cloned sequences or can be synthesized using means known in the art. PNA probes can also be used to enhance the specificity of assays.

In some embodiments, panels of nucleic acid sequences representing different regions of the genome of the microbe can be used simultaneously or sequentially to detect the microbe. In still another embodiment, panels of nucleic acid probes from different microbes can be used in the diagnostic assays described above. The probes, or oligonucleotides comprising probe sequences, can be immobilized on a substrate (e.g., a microarray) as described in Cummings et al., supra, to increase the throughput of diagnostic assays.

In one aspect, the candidate sequence, or a sequence identified by the candidate sequence, is used to express a peptide, for example, by operably linking the candidate sequence to a promoter sequence in an expression vector. In some embodiments, the candidate sequence is linked in frame to a cleavable amino acid sequence whose expression is operably linked to the promoter sequence. Alternatively, a peptide can be synthesized using the predicted amino acid sequence of the candidate sequence or a coding sequence of the sequence identified by the candidate sequence. Preferably, the peptide is an antigenic peptide. The peptide can be used to generate antibodies which specifically bind to the peptide and to polypeptides or proteins comprising the peptide.

Methods of generating antibodies are known in the art and are described in Kohler and Milstein, *Nature* 256: 495-497 (1975); Kosbor et al., *Immunology Today* 4: 72 (1983); Cote et al., *Proc. Natl. Acad. Sci. U.S.A.* 80: 2026-2030 (1983), (Morrison et al., *Proc. Natl. Acad. Sci. U.S.A.* 81: 6851-6855 (1985); Neuberger et al., *Nature* 312: 604-608 (1984); Takeda et al., *Nature* 314: 452-454 (1985); and U.S. Patent No. 4,946,778, the entireties of which are

incorporated by reference herein. Antibodies encompassed within the scope of the invention include, but are not limited to, monoclonal antibodies, polyclonal antibodies, double chain antibodies, single chain antibodies, chimeric antibodies, antibody fragments comprising at least one antigen binding site, and the like.

5 In one aspect, antibodies specific for peptides expressed by nucleic acids from the microbe are used in histological assays, such as immunohistochemistry, immunofluorescence, immunoelectron microscopy, and the like. However, antibodies can also be used in immunoassays as are routine in the art. The detection of binding of an antibody to a sample from a host organism suspected of harboring a microbe can be used to provide a diagnosis that the
10 organism harbors the microbe (e.g., that the microbe may be found on or within its cells, or in bodily fluids from the organism). For example, the antibodies according to the invention can be used to detect microbes which are shed by host cells and which may be present in bodily fluids outside of cells or in proximity to cells or tissues from the host organism, or to detect antigens which are presented after processing of polypeptides of a microbe by host cells (e.g., by host cell
15 MHC class I molecules), or to detect microbes which typically exist extracellularly within a host organism, such as bacteria.

As with the nucleic acid probes described above, panels of antibodies specific for a single microbe can be used as probes, either simultaneously or sequentially. Panels of antibodies specific for a plurality of microbes can also be used. In one embodiment, antibodies are arrayed
20 on a substrate to increase the throughput of the analysis.

In some aspects, peptides themselves can be used as diagnostic reagents. For example, peptides can be reacted from sera from an organism suspected of containing a microbe to detect the presence of circulating antibodies which react with the peptides.

Antisense Nucleic Acid Molecules

25 In one aspect, the invention provides a sequence which is a complement or an antisense sequence of a coding sequence of the candidate sequence or of the coding sequence of another sequence which has been identified by the candidate sequence. The antisense sequence can be administered to a host organism in an amount sufficient to prevent the expression of a

polypeptide encoded by the candidate sequence or the other sequence identified by the candidate sequence. Techniques of generating antisense constructs are known in the art and are described in, for example, Stein et al., *Cancer Research* 48: 2659-2668 (1988); Walder, *Genes & Development* 2: 502-504 (1988); Marcus-Sekura, *Anal. Biochemistry* 172: 289-295 (1988); Zon, *J. of Protein Chemistry* 6: 131-145 (1987); Zon, *Pharmaceutical Research* 5: 539-549 (1988); and Loose-Mitchell, *TIPS* 9: 45-47 (1988); the entireties of which are incorporated by reference. Antisense nucleic acids according to the invention additionally can be modified to enhance their stability in vivo, as described in Agrarwal et al., *Proc. Natl. Acad. Sci. USA* 85: 7079 (1988), and Sarin et al., *Proc. Natl. Acad. Sci. USA* 85: 7448 (1988), for example, the entireties of which are incorporated herein by reference.

Antisense nucleic acids also can be modified to include a cleaving agent for cleaving a molecule to which the antisense nucleic acid binds. For example, the nucleic acid can be engineered to sequences which provide the function of a ribozyme. Sequences for use in constructing ribozyme vectors are described in, for example, Rossi et al., *Aids Research and Human Retroviruses* 8: 183 (1992); Hampel and Tritz, *Biochemistry* 28: 4929 (1989); Hampel et al., *Nucleic Acids Research* 18: 299 (1990); Perrotta et al., *Biochemistry* 31: 16 (1992); Guerrier-Takada et al., *Cell* 35: 849, (1983); U.S. Pat. No. 4,987,071; Scanlon et al., *PNAS* 88: 10591-5 (1991); Dropulic et al., *J Virol.* 66: 1432-41 (1992); Weerasinghe et al., *J Virol.* 65: 5531-5534 (1991); Ojwang et al., *PNAS* 89: 10802-10806 (1992); Chen et al., *Nucleic Acids Res.* 20: 4581-1589 (1992); and Sarver et al., *Science* 247: 1222-1225 (1992); the entireties of which are incorporated herein by reference.

Antisense molecules can be administered directly to a target site. For example, antisense molecules can be administered topically (e.g., to skin), by direct injection into cells (e.g., such as tumor cells), by direct administration to a tissue which has been exposed by surgery, or through a medical access device, such as a catheter or endoscope, which can deliver the molecule directly to the target site (e.g., by bringing the tissue into contact with a solution comprising the antisense molecules). In another aspect, antisense molecules are administered to the patient enterally or parenterally. Antisense molecules can be administered with suitable carrier molecules to facilitate delivery to a target site (e.g., by complexing the molecules with liposomes) and/or can be bound to a targeting molecule (e.g., a ligand specific for a receptor expressed on the surface of

a host cell infected by the microbe). Preferably, the targeting molecule includes an intracellular localization signal for delivering the antisense molecule to the interior of the cell.

Therapeutic Peptides

As discussed above, candidate sequences, or sequences identified by these sequences, can be used to generate peptides. In some aspects, the peptides are administered to the host organism in an amount sufficient to enable the host organism to mount a protective immune response against the microbe. In a preferred embodiment, the peptides are used as a vaccine.

Alternatively, or additionally, nucleic acid sequences which encode these peptides and which are operably linked to one or more promoter elements can be administered to the host organism in an amount sufficient to enable the host organism to mount a protective immune response against the microbe (e.g., providing a DNA vaccine). A protective immune response can include the production of macrophages which specifically recognize the microbe (e.g., during an extracellular portion of its life cycle) and/or the production of cells which produce neutralizing antibodies which specifically bind to the microbe and which prevent the microbe from infecting further cells.

In some aspects, a plurality of peptides from the same microbe or a nucleic acid expressing the plurality of peptides is administered to the organism. In some embodiments, the microbe is isolated and nucleic acids removed, and the microbe itself is administered to an organism to generate a protective immune response (see, e.g., as described in U.S. Patent No. 5,698, 430, the entirety of which is incorporated by reference herein).

Examples

The invention will now be further illustrated with reference to the following examples. It will be appreciated that what follows is by way of example only and that modifications to detail may be made while still falling within the scope of the invention.

25 Example 1

Computational subtraction was used to identify sequences in an EST library (Unigene library #271) from the HeLa cervical carcinoma cell line. This library contains 7,073 EST's.

6,752 of these EST's comprise at least 100 discrete, unambiguous 15-mers (e.g., sequences whose nucleotide identity can be assigned at greater than 95% confidence levels or 0% N's). A system 1 according to the invention was used to compare the sequences in the EST library against known human mRNA sequences, human repeat sequences, human mitochondrial sequences, the Human Genome Project (HGP) and Celera Genomics Human Genomic DNA sequences and to eliminate matching sequences. Matches within mouse genomic DNA sequences (Celera) were also searched for and removed under the assumption that these would represent unsequenced regions of the human genome.

Using a BLAST score cut-off of 60, equivalent to 30 consecutive identical nucleotides, and an expectation value of 10^{-7} as a cutoff, the 7,073 EST's were pared down to 144 non-matching sequences. Application of a quality filter to set a minimum length cut-off of 150 nucleotides and a 2.5% maximum ratio of ambiguous nucleotides (e.g., > 2.5% N's) to non-ambiguous nucleotides, decreased the number of matching sequences to 43. When the 43 remaining sequences were matched to the GenBank nt database, 17 sequences matched additional human mRNA sequences and 6 matched known E. coli sequences.

Thus, using the system 1, 7,073 EST's in HeLa cells were reduced to a total of 22 sequences that failed to match human, mouse, or E. coli genomic sequences. Two of these sequences were subsequently determined to be identical to human papillomavirus (HPV) type 18 sequences. HPV is a cause of cervical cancer and HPV nucleic acids are known to be present in the HeLa cell line (see, e.g., Boshart et al., 1984, EMBO 3(5): 1151-1157). Two other HPV sequences were present in the HeLa cell EST library, but were filtered out by the system 1 because of a match to sequences in the Celera genome database. These two sequences match a HPV type 45 sequence from the NCBI database that was included in the Celera genome assembly but not in the NCBI assembly, thus, verifying the ability of the system 1 to identify microbial sequences through computational subtraction.

To determine which of the unmatched sequences (i.e., candidate sequences) represented unsequenced human genomic DNA and to determine whether candidate sequences could be used to identify pathogenic DNA (e.g., such as HPV DNA), PCR primers corresponding to each of the 22 non-matching sequences were generated and tested on a panel of normal human genomic

DNA samples and HeLa cell genomic DNA. Ten primer sequences were capable of amplifying nucleic acids in all samples of human genomic DNA, while ten primer sequences could not amplify any samples, and two primers (corresponding to HPV sequences) were able to amplify only HeLa cell genomic DNA. The ten sequences which amplified all human genomic DNA samples are likely to represent previously unsequenced regions of the human genome, while those primer sequences unable to amplify sequences in any samples are likely to represent sequences brought together by splicing (and which are therefore too far apart in genomic DNA to be amplified), sequences of non-human origin, or sequencing errors.

These results demonstrate that the system 1 is capable of identifying microbial sequences (e.g., such as HPV sequences) by computational subtraction.

Example 2

Given the ability of computational subtraction to detect viral sequences in HeLa cells, computational subtraction was used to scan existing EST databases for candidate microbial sequences using the same method as described in example 1. EST's in a NCBI EST database of 3,287,578 sequences were serially compared against filter databases using the MEGABLAST tool with a word-size of 24, to filter or subtract matching sequences. The sequences were filtered through a known human mRNA database (the NCBI RefSeq human mRNA database), which after subtraction left 1,438,967 sequences, a human mitochondria database, which after subtraction left 1,409,118 sequences, a vector sequence database (the NCBI UniVec database), which after subtraction left 1,396,697 sequences, a human repetitive sequence database (GIRST HumRep), which after subtraction left 1,368,895 sequences, a human genome database, which after subtraction left 144,498 sequences, and a mouse genome database, which after subtraction left 137,011 sequences. To improve the sensitivity of the filtering process, remaining EST's were re-run against the filters at a lower word size (20) and matching EST's were again removed, leaving 120,792 EST's unmatched. The process was repeated using a word size of 16, leaving 102,009 EST's unmatched. This last sequence set was passed through a quality filter as described in Example 1, to remove short (< 150 nucleotides) and ambiguous (> 2.5% N's) sequences. At the end of this subtraction or filtering process, 65,839 sequences of 3,287,578

EST sequences or 2% of the EST sequences in the NCBI EST database were found not to match a panel of human genomic or reference sequences.

Sequences were subsequently tested by BLASTN searches against GenBank nt databases (i.e., a database comprising multiple species' sequences, including microbial sequences) using
5 (using a word size of 16) and by BLASTX searches against the nr non-redundant protein databases (using a word size of 3). A results database of these matches is available at <http://www.hcs.harvard.edu/~weber/meyerson2/nrnt.cgi>, the entirety of which is incorporated by reference herein.

Despite filtering sequences against human genomic and other databases (e.g., removing
10 matching sequences), a significant fraction of the remaining EST's still matched nucleotide and/or protein sequences of known human origin. In total, 5,119 "non-matching" candidate sequences matched nucleotide sequences from Homo sapiens using a BLASTN score minimum of 100, while the translations of 211 sequences, without nucleotide matches, matched Homo sapiens protein sequences, with a BLASTX minimum score of 100. These data are consistent
15 with the as-yet incomplete sequencing of the human genome.

Strikingly, a significant number of sequences with matches to viral, fungal, bacterial, and plant sequences were found in the non-matching, i.e., candidate sequence set. A culled set of matching species sequences was generated by excluding all vertebrate, as well as Escherichia, Saccharomyces, Drosophila, and Caenorhabditis sequences that might represent library
20 contamination. Using BLASTN and BLASTX minimum scores of 100, 1,055 sequences were found which matched nucleotide sequences from the culled species (i.e., sequences representing likely contaminants) and 759 sequences were found which matched culled protein sequences but not nucleotide sequences. Matches to microbial sequences are described in Tables 1 through 3.

Table 1. Viral Genomes With Nucleotide Similarity to Filtered Human EST Sequences				
Viral Species	EST Count	Library Count*	Tissue Types	Most Common Library No.***
Hepatitis B virus	33	2	adult liver, hepatocellular carcinoma	3618 (adult liver)
Human spumaretrovirus	10	1	fetal liver	168 (fetal liver)
Cytomegalovirus	9	3	nervous system, breast, uterus	2915 (nervous system)
Human adenovirus 2	8	6	lymph, ovary (2)**, colon, head and neck, lung	2222 (lymph)
Simian sarcoma virus	7	6	breast	3633 (breast)
Human papillomavirus (subtypes 16 and 18)	7	3	cervix, placenta, uterus, tumor	271 (cervical carcinoma cell line, i.e., HeLa)
Stealth virus 1	4	2	head and neck (2)	4582 (head and neck)
Kaposi's sarcoma associated virus (HHV-8)	3	2	nervous, head & neck	2836 (nervous system)
Hepatitis C virus	2	1	bone marrow	4862 (bone marrow)
Epstein-Barr virus (HHV-4)	1	1	lymph	5167 (lymph)
<p>* "Library Count" reflects the number of libraries in which EST matches to a particular virus were found.</p> <p>** The total number of different libraries of a given tissue type is indicated in parentheses (if greater than one).</p> <p>*** Library numbers are based on UniGene assignments.</p>				

EST sequences that passed all filters (e.g., remained present after computational subtraction against one or more databases) were compared to GenBank's nt database (a database representing multiple species) using the MEGABLAST algorithm. Alignments with a bit score of 100 or greater were categorized as "matching" those in the nt database. Sequences remaining

after subtraction which match viral genome sequences are shown in Table 1. Included in these sequences were sequences belonging to a variety of pathogenic viruses. As shown in Table 1, the most common viral match was to Hepatitis B virus sequences, for which there were 33 EST matches in the databases. Thirty-two of these matches were derived from the library GKC which is made from normal liver tissue from a Chinese patient with hepatocellular carcinoma. Hepatitis B virus sequences are abundant in this library, representing 0.2% of the 16,743 total sequences in this library. As seen in Table 1, a variety of other pathogenic virus sequences including human papillomavirus; adenovirus; and a variety of herpesviruses, including cytomegalovirus, Epstein-Barr virus, and Kaposi's sarcoma herpesvirus; were identified by computational subtraction methods according to the invention.

Table 2, below, summarizes sequences remaining after computational subtraction which match bacterial sequences. After identifying expressed sequences as candidate sequences not found in the human genome, these sequences were compared to the GenBank nt database using the BLASTX algorithm (BLAST 2.0) and alignments with a bit score of 100 or greater were categorized as matches. Table 2 shows the ten most frequently appearing bacterial sequences after computational subtraction. As can be seen from Table 2, there are numerous matches to *Pseudomonas aeruginosa* sequences, a common pathogen as well as a commensal organism. In addition, there are numerous matches to other *Pseudomonas* species.

Bacterial species	EST Count	Library Count*	Tissue Types	Most Common Library No.***
<i>Pseudomonas aeruginosa</i>	304	85	breast (21)**, head and neck (11), bone marrow (8)	3025 (bone marrow)
<i>Xylella fastidiosa</i>	92	32	head and neck (12), breast (10), stomach (3)	1304 (breast)
<i>Pseudomonas</i> sp.	56	11	breast (3), lymph 92), B-cells, muscle, ovary	4873 (uterus)
<i>Pseudomonas putida</i>	32	17	head and neck (6), breast (5), bone marrow (2)	1148 (breast)
<i>Caulobacter crescentus</i>	29	7	thymus (2), colon, lymph, uterus	3587 (thymus)

Table 2 (cont'd). Bacterial Genomes With Nucleotide Similarity To Filtered Human EST Sequences				
Mesorhizobium loti	26	13	lymph (2), thymus (2), foreskin, breast	2223 (lymph)
Fusobacterium naviforme	17	5	head and neck (4), uterus	4796 (head and neck)
Leptotrichia-like sp.	17	10	uterus (9), head and neck (1)	4796 (uterus)
* "Library Count" reflects the number of libraries in which EST matches to a particular bacteria were found.				
** The total number of different libraries of a given tissue type is indicated in parentheses (if greater than one).				
*** Library numbers are based on UniGene assignments.				

The more interesting category of bacterial matches is shown in Table 3 which shows the set of bacterial sequences whose conceptual translations match known bacterial proteins and which do not share significant nucleotide sequence similarity with known bacterial nucleotide sequences. These sequences were identified by passing EST sequences through the filter databases described above and comparing remaining sequences to the GenBank nt database using the BLASTN algorithm (with a threshold of 60 bits) and to the non-redundant ("nr") protein database using the BLASTX algorithm (setting a threshold of 100 bits). EST's matching the nr database but not the nt database were categorized as "translation-only alignments." These series of operations revealed numerous pathogens with matches only to translated sequences. Again, many matches were found to *Pseudomonas aeruginosa* sequences. Other candidate sequences included those whose translated sequences matched proteins of *Mycobacterium tuberculosis*, *Vibrio cholerae* and *Neisseria meningitidis*. This suggests the presence of clones representing novel unsequenced bacteria, highly related to these pathogens, but previously undescribed, in the libraries.

Table 3. Sequences in Human EST Libraries With Translation Matches to Bacterial Sequences

Bacterial Species*	EST count	Library Count**	Most Common Tissue Types	Most Common Library No. ****
<i>Pseudomonas aeruginosa</i>	239	96	breast (20)***, head and neck (14), bone marrow (8), CNS (7)	3587 (thymus)
<i>Caulobacter crescentus</i>	32	17	thymus (3), breast (3), lymph (2)	2223 (lymph)
<i>Xylella fastidiosa</i>	22	14	lymph (3), thymus (2), lung (2)	3587 (thymus)
<i>Mycobacterium tuberculosis</i>	14	8	thymus (2)	2223 (lymph)
<i>Streptomyces coelicolor</i>	29	14	thymus (2), breast (2), lymph (2)	2223 (lymph)
<i>Vibrio cholerae</i>	13	11	ovary (2)	2217 (B cells)
<i>Bacillus subtilus</i>	14	10	heart (2)	47 (heart)
<i>Neisseria meningitidis</i>	12	10	thymus (3), breast (2)	650 (pooled)
<i>Pseudomonas putida</i>	11	4	bone marrow (2), lymph (1), thymus (1)	3587 (thymus)

* Bacterial species to which candidate sequences are related, as determined by matching conceptual protein translations of candidate sequences but not matching nucleotide sequences of candidate sequences.

** "Library Count" reflects the number of libraries in which EST matches to a particular bacteria were found.

*** The total number of different libraries of a given tissue type is indicated in parentheses (if greater than one).

**** Library numbers are based on UniGene assignments.

In each of the Examples discussed above, sequence analysis was performed using sequence available from the NCBI (<http://ncbi.nlm.nih.gov>), Celera Genomics (<http://www.celera.com>) and the Genetic Information Research Institute (GIRI) (<http://www.girinst.org>). Human EST sequences and library information, Human Genome Project Sequences (phases 0-3), the RefSeq human mRNA set, and UniVec vector sequences

were downloaded from NCBI on March 6, 2001. The "nr" and "nt" BLAST databases were downloaded from NCBI on March 30, 2001. The human mitochondrial genome sequence is GenBank accession # NC_001807. The HeLa cell EST library analyzed is available as Library 271 (Stratagene_HeLa_cell_s3_937216) in the UniGene resource at the NCBI web-site. The
5 Celera draft of the human genome and the 3x coverage of shotgun sequence from the mouse genome were downloaded from Celera's website in January, 2001. RepBase6.2 was downloaded from the GIRI database on March 7, 2001.

Variations, modifications, and other implementations of what is described herein will occur to those of ordinary skill in the art without departing from the spirit and scope of the
10 invention.

What is claimed is:

11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2